The use of PacBio and Hi-C data in denovo assembly of the goat genome

Presented by Derek Bickhart
Work performed by Sergey Koren, Adam Phillippy, Tim Smith, Josh Burton, Ivan Liachko, Tad Sonstegard, Curt Van Tassell, Heather Huson, and Steven Schroeder

Existing Goat Genome

- CHI 1.0
  - 542,145 contigs
  - 18.7 kb contig N50
- Placed chromosomes
  - Based on cattle genome comparison
  - 30 “pseudo-chromosomes”

Selecting an animal for sequencing

- Illumina Goat SNP chip to screen 4 U.S. goat breeds
- Average inbreeding coefficient in San Clemente breed above 0.40
- The San Clemente breed is a “rescued” feral population
  - Two severe bottlenecks since 1875
  - Papadum has genomic inbreeding co-efficient of 0.42 and a Y-Chr

Genome Assembly: Generating Contigs

- Contig: Contiguous sequence
- Consensus overlap (String graph)
- De Bruijin graph-based

The fundamental problem: repeats

- Read length > Repeat size
- Larger repeats
Sequencing technologies suited to assembly

<table>
<thead>
<tr>
<th>Max Read Length</th>
<th>Illumina</th>
<th>Sanger</th>
<th>454</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>250</td>
<td>750</td>
<td>1,000</td>
</tr>
</tbody>
</table>

• PacBio error rate is random
• High coverage, consensus correction

Starting dataset

• Initial filtered dataset

| Total Number of Sequenced Bases | 194 Gbp |
| Total Number of Reads | 38 Million |
| Mean read size (bp) | 5,110 |
| Maximum read size (bp) | 40,530 |

PacBio assembly creates incredible contigs

| Total bases | 2.59 Gbp |
| Number of Contigs | 5,902 |

PacBio identified inversion

More Sequence in the PacBio assembly

Major Structural Inconsistencies

From Chen et al. 2013

Tim Smith
Sergey Koren
Adam Phillippy

Sergey Koren
Adam Phillippy
Genotyping platform correction

Trans-contig error:

![Diagram showing trans-contig error]

Contig distance error:

![Diagram showing contig distance error]

Practical results gleaned from this assembly

- Initial relocation of SNP markers
  - Trans-contig errors: 8 SNP pairs
  - Contig distance errors: 16 SNP pairs

- Chromosome X probe sequence
  - Original regions checked for 120 probes
  - Probe sequence was chimeric
  - 100% had new assembly corrections

Future directions

- Scaffold into Karyotype chromosomes
  - Hi-C data (Lachesis)
  - Radiation Hybrid map

- Provide correct SNP marker placement

- Better resources for the community

Acknowledgements

- USDA ARS
  - Tad Sonstegard
  - Curt Van Tassel
  - Tim Smith

- NBACC
  - Adam Phillippy
  - Sergey Koren

- Jay Shendure Lab
  - Ivan Liachko
  - Josh Burton

- Virginia University
  - Brian Sayre

- Cornell
  - Heather Huson

- NIFA/AFRI

- USDA USAID – Feed the future.